Közösségkeresés alapú felügyelet nélküli szófaji egyértelműsítés

Berend Gábor[1], Vincze Veronika[2]

[1]Szegedi Tudományegyetem, TTIK, Informatikai Tanszékcsoport,
Szeged, Árpád tér 2., e-mail:berendg@inf.u-szeged.hu
[2]Magyar Tudományos Akadémia, Mesterséges Intelligencia Kutatócsoport,
Szeged, Tisza Lajos körút 103., e-mail:vinczev@inf.u-szeged.hu

Kivonat Az előadásban bemutatjuk felügyelet nélküli szófaji egyértelműsítő módszerünket, mely közösségkeresésre épül. A közösségkereső eljárás bemenetéül szolgáló, a szóalakok fölött értelmezett hasonlósági gráf költséges számítására való tekintettel az elosztott rendszerek területén az ún. overlay topológiák közelítésére korábban már sikeresen alkalmazott T-MAN algoritmust alkalmaztuk. Eredményeink azt igazolják, hogy sikerült átültetnünk a két különböző tudományos közösség által használt módszerek előnyeit a szófaji egyértelműsítés területére, azaz egy olyan feladatra nyújtottunk így megoldást, amelyet egy harmadik tudományos közösség tűzött ki céljául.

Kulcsszavak: szófaji egyértelműsítés, közösségkeresés, felügyelet nélküli tanulás, modularitás

1. Bevezetés

A szófaji egyértelműsítés a természetes nyelvi feldolgozás egyik alapvető lépése: számos magasabb rendű alkalmazás hasznosítja jellemzőként a szófaji kódokat, azaz igen fontos, hogy a szövegszavakhoz hozzárendeljük azok szófaji elemzését. A felügyelt szófaji egyértelműsítési módszerek nagyméretű, kézzel annotált adatbázisokra épülnek. Az annotált adatbázis létrehozásához azonban szükséges egy, az adott nyelvre kidolgozott morfológiai kódrendszer is, melynek segítségével morfológiailag elemezni és egyértelműsíteni lehet az adott nyelvű szövegeket. Bizonyos nyelvekre azonban nem áll rendelkezésre ilyen kódrendszer és/vagy nagyméretű annotált adatbázis. Ez esetekben a megoldást a félig felügyelt vagy felügyelet nélküli szófaji egyértelműsítési módszerek jelenthetik, melyek segítségével az ilyen nyelvekre is lehetséges hatékony szófaji egyértelműsítőt építeni.

A felügyelt szófaji egyértelműsítési módszerek a szövegszavakat előre meghatározott (a tanító adatbázisban szereplő) szóosztályokba sorolják. Azonban előfordulhat, hogy egy nyelvre többféle annotációs rendszer is létezik, más-más mennyiségű elérhető annotált adattal, ami megnehezíti a különféle szófaji egyértelműsítő módszerek hatékonyságának összevetését. Például a hunpos tagger [1]

a KR morfológiai kódrendszerre épül, ám jelenleg nem tudunk olyan kézzel annotált adatbázisról, amely a KR-kódokat használná. Így a hunpos hatékonyságát csak úgy lehetséges mérni, ha a KR-kódokat megfeleltetjük egy kézzel annotált korpuszban szereplő kódoknak, ami szintén idő- és munkaigényes feladat.

A felügyelet nélküli szófaji egyértelműsítő módszerek különféle csoportokba (klaszterekbe) sorolják a szavakat, így képesek kiküszöbölni a fenti hátrányokat, mivel a klaszterek összevethetők bármely morfológiai kódrendszer által alkalmazott csoportokkal. A módszer tovább előnye, hogy a szófaji egyértelműsítés részletességét különböző technikákkal lehetséges szabályozni. Míg egyes kódrendszerek túlságosan részletes kódokat tartalmaznak (például képzéssel kapcsolatos információkat), addig a legtöbb alkalmazás számára nem szükséges a kódok ilyen mértékű részletezése: a fő szófaj megadása általában elégségesnek bizonyul a legtöbb alkalmazás számára (például információ-visszakeresés, névelemfelismerés vagy kulcsszókinyerés). Ezzel szemben más esetekben fontos lehet a minél részletesebb morfológiai információ, például a gépi fordításban vagy a szemantikai szerepek meghatározásában a főnévi esetragok igen nagy szereppel bírnak. A szükséges részletességet a klaszterek mennyiségének befolyásolásával tudjuk biztosítani. Az aktuális feladat számára indokolt klaszterszám befolyásolására a T-MAN [2] hálózati topológiaépítő pletykaalgoritmus számára bemenetként adott gráf eltérő módokon történő felépítésével nyílik lehetőség.

Az általunk használt közösségkereső eljárás [3] a szóalakok kontextuális tulajdonságaiból épített hálózat particionálásával állítja elő az egyes lexikai csoportokat. A gráfelméleti alapokon nyugvó algoritmus a particionálandó gráfok legjobb modularitással járó felbontására ad kielégítő és gyors közelítést. Az eljárás egy további tulajdonsága, hogy mivel a különböző particionálásokat jellemző modularitás mérőszámának több lépésben végrehajtott maximalizálásával történik, így lehetőség van hierarchikus közösségek kialakítására, amelyek a felhasználási területtől függően eltérő hasznossággal bírhatnak, hiszen a szóalakok durvább és részletezettebb lexikai csoportokba sorolása is lehetséges.

Eredményeink azt igazolják, hogy megközelítésünk felveszi a versenyt az angolra alkalmazott felügyelet nélküli módszerekkel, mindemellett a módszer magyarra való alkalmazhatóságát is számszerűsítettük.

2. Kapcsolódó munkák

A felügyelet nélküli és félig felügyelt szófaji egyértelműsítés területén már számos korábbi munka született az utóbbi évtizedekben, melyek több csoportba sorolhatók. Az egyik megközelítés szerint a kívánt szófaji klaszterek számát előre meg kell adni [4,5], ugyanakkor más rendszerek a klaszterek számát az adott feladathoz igazítva határozzák meg. Míg egyes módszerek rejtett Markov-modellekre épülő felügyelet nélküli tanulásként tekintenek a problémára [6,7], addig mások magasabb dimenziós terekben végeznek számításokat, illetve megint mások gráfként közelítenek a problémához. Továbbá, bizonyos módszerek működéséhez szükség van egy előre megadott részleges szótárra vagy néhány mintapéldára is, azonban ezek nem minden esetben állnak rendelkezésre.

Számos kiértékelési metrika használatos a szakirodalomban, melyek gyakran a több szófaji klasztert előállító módszereket részesítik előnyben. A legtöbb szerző azonban az információelméletből kölcsönzött V-mérték mellett teszi le a voksát [8]. A felügyelet nélküli szófaji egyértelműsítő módszerek kiértékelése megfeleltetés alapján is történhet, amikor is a rendszer teljesítményét a létrejott klaszterek (vagy ezek egy részhalmaza) és az etalon klaszterek közti megfeleltethetőség alapján határozzák meg. A kiértékelési metrikákról [9] ír bővebben.

A hálózatelemzés kulcsfontosságú szereppel bír a felügyelet nélküli megközelítésekben, ahol a magasabb dimenziós terekben történő klaszterezés helyett gráfalapon hajtódik végre a művelet, figyelmen kívül hagyva a dimenzionalitást. A hálózatelemzési módszerek közül különösen a közösségkeresés kapott nagy figyelmet több tudományterületen is a biológiától kezdve a szociológián át az informatikáig. A gráfok particionálása kapcsán a modularitás vált meghatározó fogalommá a korábbi metrikák közül [10]. A modularitás eredetileg a gráf particionálásának hatékonyságát hivatott mérni, és később számos gráfparticionáló algoritmus – mint például a spektrális optimalizáció, mohó algoritmusok és szimulált hűtés – célfüggvényévé vált.

3. Módszertan

A közösségkereső eljárásra épülő szófaji egyértelműsítés az eltérő szóalakok fölött értelmezett hasonlósági gráf particionálásán alapul, amely hasonlósági gráf építésének és jellemző csoportokra bontásának részletes bemutatására a következőkben kerül sor.

3.1. Hasonlósági gráf

Mivel a hasonló kontextusban szereplő szóalakokról feltételezhető, hogy hasonló mondatbéli funkcióval is bírnak [11], ezért eljárásunkban a szóalakok szófaji kategóriáinak felügyelet nélküli meghatározására egy olyan eljárást valósítottunk meg, mely a szóalakok fölött értelmezett hasonlósági gráf particionálásán alapul. Algoritmusunk a szóalakokat a hozzájuk meghatározott kontextusvektorok alapján sorolja be a hasonló szerepet betöltő és általunk azonos szófajúként interpretált szavak halmazaiba. Első lépésként tehát a szóalakok fölött értelmezett, súlyozott hasonlósági gráfunkat definiáljuk.

Munkánk során a szófajuk szempontjából csoportosítandó szavak alkották azt a V szótárat, amely elemeit eltérő méretű ($1 \leq W \leq 3$) ablakok mellett vett szókörnyezet-eloszlásokkal jellemeztük. (Mind a csoportosítandó szóalakok meghatározása során, mind pedig a környezetük vizsgálata során egy egyszerű reguláris kifejezés segítségével a numerikus kifejezéseket egységesen kezeltük.) A különböző méretű és nyelvű korpuszok feldolgozása során egy-egy szóalakot, a bal és jobb oldalukon, eltérő $w \leq W$ pozíciókon számított $2*(|V|+1)*W$ méretű eloszlásvektorral jellemeztünk. A későbbiekben particionálandó hasonlósági gráf csúcsait a $|V|$ méretű szótár egy-egy eleme képezte, a csúcsok közötti élsúlyok

meghatározásában pedig a szóalakokhoz társított eloszlásvektorok játszottak szerepet.

A gráfalapú megközelítések előnye többek között az, hogy a kiugró értékek (outliers) kezelése viszonylag természetes módon kezelhető szemben például a k-közép klaszterezéssel. A nem releváns és így nem kívánt hasonlóságok kiszűrésének egy lehetséges módja a teljes gráfokról a k-legközelebbi gráfokra való áttérés lehet. Azon túl, hogy a gráfban csökkenthető a zajt okozó kapcsolatok száma, a gráf ritkításával egyúttal jótékonyan befolyásolható a gráfon végzett algoritmusok sebessége.

Éppen ezért a szóalakok egymáshoz való viszonyának reprezentálása során a teljes gráfokból $G_k = (V, E_k, w)$ k-legközelebbi szomszédságon alapuló gráfokat konstruáltunk, melyekre $E_k = \{(u, v) : n(u, k) \ni v \vee n(v, k) \ni u\}$, ahol az $n(u, k)$ és $n(v, k)$ függvények rendre az u és v csúcsokhoz tartozó k legközelebbi szomszédot adják vissza, $w(u, v)$ pedig az u és v csúcsok közötti szimmetrikus távolságot határozza meg. A csúcsok közötti távolságot a *koszinusz távolság* (1), *Jensen-Shannon divergencia* (2), illetve *Jaccard-együttható* (3) segítségével is vizsgáltuk, melyek kiszámítása a következő képletek alapján történt:

$$cos(q, r) = 1 - \frac{\sum_v q(v)r(v)}{\sqrt{\sum_v q(v)^2}\sqrt{\sum_v r(v)^2}} \tag{1}$$

$$JS(q, r) = \frac{1}{2}[D(q\|avg_{q,r}) + D(r\|avg_{q,r})] \tag{2}$$

$$jacc(q, r) = 1 - \frac{|\{v : q(v) > 0 \wedge r(v) > 0\}|}{|\{v|q(v) > 0 \vee r(v) > 0\}|} \tag{3}$$

Az előzőekben bemutatott metrikák valamelyikével a csúcsokhoz történő k legközelebbi szomszéd meghatározását követően az eddig távolságokként értelmezhető élsúlyokat hasonlósági értékekké alakítottuk át. A hasonlósági mértékre való áttérés érdekében minden (u, v) csúcs közötti súlyt a $sim(f(u, v)) = \frac{1}{1+f(u,v)}$ képletnek megfelelően alakítottuk át, ahol $f(u, v)$ az előzőekben definiált távolságfüggvények értéke u és v csúcsokra nézve. A távolság helyett a hasonlósági értékekre való áttérésnek a közösségkereső eljárás súlyozott gráfon értelmezett működése kapcsán volt fontos.

3.2. Modularitásalapú közösségkeresés

Az általunk használt, modularitás maximalizálására építő eljárás előnye, hogy a kialakuló közösségek száma a particionálandó gráf topológiája alapján kerül meghatározásra, szemben egyéb eljárásokkal (pl. k-közép klaszterezés). Egy adott gráfparticionálást jellemző modularitás kiszámításával egy jósági értéket rendelhetünk a felbontás minőségére nézve, mely figyelembe veszi a gráf topológiájából adódóan az egyes csúcspárok között elvárható élek számát, valamint egy tényleges felbontás során az egyes csoportokon belül vezető élek tapasztalt számát. Az

előzőekben elmondottak a következő képlettel számolhatók:

$$Q = \frac{1}{2m} \sum_{ij} (A_{ij} - \frac{k_i k_j}{2m}) \delta(C_i, C_j) \tag{4}$$

, amelyben az összegzés minden *lehetséges* élre (minden i és j csúcsra) vonatkozik, és ahol az A_{ij} a particionálandó gráf szomszédsági mátrixának egy eleme, m a gráfban található élek száma, az összegzésben található hányados az i és j csúcsok között menő élek várható értéke, a δ függvény pedig az ún. Kronecker-delta, mely akkor veszi fel az 1 értéket, ha az i és a j csúcsok megegyező klaszterben találhatók, máskülönben 0.

Számos jó tulajdonsága miatt vonzó elgondolás lenne a gráfokhoz olyan felbontásokat keresni, amelyek a modularitás jósági mérőszámát tekintenék célfüggvényül, azt maximalizálnák. Ugyanakkor ahogy arra már rámutattak [12], ez a feladat erősen $\mathcal{NP}$-teljes. A negatív eredményből adódóan, számos közelítő eljárás látott napvilágot a probléma kezelhető időben történő minél hatékonyabb megoldására, melyek között találunk szimulált hűtéstől kezdődően spektrálmódszereken át mohó megközelítéseket is.

Ugyan a spektrálmódszereken alapuló eljárások gyakorta jobb eredményeket érnek el más megközelítésekhez képest, nagyméretű gráfok esetében sokszor nem hatékonyak, és mivel esetünkben kifejezetten nagy gráfok felbontását kíséreltük meg, így kiemelten fontos volt, hogy a maximális modularitást eredményező felbontás közelítésére alkalmazott eljárásunk számítási igénye alacsony legyen. A [3] által alkalmazott mohó optimalizáló stratégia kifejezetten nagy gráfokon is működőképesnek bizonyult, így az általuk javasolt eljárást valósítottuk meg a szóalakok gráfjának maximális modularitást elérő felosztásának meghatározására. A szerzők által javasolt eljárás egy alulról felfelé építkező klaszterező eljárás, mely kezdetén minden csúcsot egy külön klaszterbe sorolnak, majd a további lépések során a csúcsok meglátogatása során azokat a lokálisan legjobb modularitás növekményt eredményező közösséghez sorolják (esetleg egyikhez sem). Egy i csúcs C közösségbe történő mozgatása során kettős hatás figyelhető meg: egyrészt növeli a globális modularitás értékét azon élei által, amelyek immáron a C közösségbeli szomszédjaival való összeköttetést biztosítják, másrészről viszont a modularitás bizonyos mértékű csökkenése is megfigyelhető lesz azon élei kapcsán, amelyek a korábbi közösségének tagjaival való összeköttetésért voltak felelősek. Egy i csúcs C közösségbe történő átmozgatásának hatása a következők szerint összegezhető:

$$\Delta Q = \left[\frac{\sum_{in} + k_{i,in}}{2m} - \left(\frac{\sum_{tot} + k_i}{2m} \right)^2 \right] - \left[\frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m} \right)^2 - \left(\frac{k_i}{2m} \right)^2 \right] \tag{5}$$

, ahol $\sum_{in}$ és $\sum_{tot}$ értékek rendre a C közösségen belül, illetve a C közösséget érintő élek súlyainak összege, k_i és $k_{i,in}$ pedig rendre az i csúcsot tartalmazó, illetve az i csúcsot a C közösséggel összekötő élek súlyainak összege, m pedig a particionálandó gráfban található élek összsúlya. Miután minden csúcs besorolást

nyert az egyes közösségekbe, az algoritmus a kialakult közösségeket összevonva, és azokat egy csúcsként kezelve megismétli az előző eljárást. Egy soron következő iterációs blokk kezdetén tehát éppen annyi csúcsot tartalmazó gráfot bontunk ismét közösségekre, amennyit az előző blokkban azonosítottunk (a korábbi blokk közösségeinek megfeleltethető élsúlyok pedig a megelőző lépésben a két közösség közt menő élek összsúlyával lesz egyenlő, a közösségen belüli élek pedig hurokélként jelentkeznek.) Az iterációs blokkokat ismételhetjük fix lépésszámig, vagy addig, amíg a modularitás növekedése fenntartható. Az eljárás előnye, hogy az eredeti hasonlósági gráf csúcsai fokszámának várható értékének fix voltából adódóan az eljáráshoz elvégzendő műveletek száma nagyságrendileg a hasonlósági gráf csúcsainak lineáris függvénye lesz. További előny, hogy az iterációs blokkok mentén eltérő finomságú – de ugyanúgy a modularitás maximalizálására törekvő – felbontásait nyerhetjük ki a particionálandó gráfnak.

3.3. A legközelebbi szomszéd gráf pletykaalgoritmussal történő közelítése

Más felügyelet nélküli módszerhez hasonlóan az általunk javasolt eljárás is nagy elemszámú minta alapján próbálja a szóalakok közt fennálló szabályszerűségeket megragadni, ami azzal jár, hogy a szótár méretének növekedésével együtt a hasonlósági gráf csúcsainak száma több százezres nagyságrendben is mozoghat, ami pedig – nagyobb W kontextusablak választása esetén – akár az egyes szóalakokat leíró szókörnyezeteloszlás-vektorok milliós hosszát is eredményezheti. Jóllehet a szókörnyezeteloszlás-vektorok jellemzően igen ritkák, egy adott esetben több százezer csúcsot tartalmazó hasonlósági gráfra még így sem határozható meg igazán hatékonyan minden szögponthoz annak k legközelebbi szomszédja.

A szótárméret növekedésével együtt jelentkező hatékonysági probléma megoldására a T-Man [2] pletykaalapú peer-to-peer protokollt hívtuk segítségül, melynek eredeti célja speciális, dinamikusan változó, nagyméretű ún. overlay hálózatok topológiájának feltérképezése. Az overlay hálózatok dinamikusságából adódóan az algoritmus a hálózati topológia egy közelítését határozza csupán meg, amire esetünkben a szóalakok hasonlósági gráfjának statikusságából adódóan ugyan nem lenne szükség, ugyanakkor a szótár méretének növekedéséből adódó problémákra megoldást nyújthat sebességével. A protokoll a következők szerint jár el: minden csúcs (peer) inicializálásra kerül egy fix méretű random szomszédos csúcsokat (peereket) tartalmazó bufferrel, majd az egyes iterációk során a csúcsok (peerek) 'kommunikálnak' egymással, amely során lehetőségük nyílik a hozzájuk tartozó bufferek tartalmának frissítésére, amennyiben azzal javítani tudnak annak tartalmán. (Esetünkben az overlay hálózatok azon speciális tulajdonságával, hogy a csúcsok folyamatosan be,- illetve kiléphetnek a hálózatból, nem kellett számoljunk.)

A szerzők algoritmusuk gyors konvergenciájáról számoltak be, vizsgálataik alapján 10-15 iteráció elégségesnek bizonyult az eredeti hálózatok topológiájának közel tökéletes közelítésére. A szóalakok fölötti hasonlósági gráf k-legközelebbi szomszédságának feltérképezése kapcsán tapasztalható konvergenciával kapcsolatos eredményeinket a 4. fejezet tartalmazza.

4. Eredmények

Az előzőekben bemutatottak szerint működő közösségkeresésen alapuló szófaji egyértelműsítőt – annak felügyelet nélküli voltából adódóan – módosítások nélkül alkalmazhattuk magyar, illetőleg angol nyelvű szövegekre. Angol nyelvű vizsgálódásaink tárgyát az ACL/DCI korpuszban található Wall Street Journal 1987. évadának 1-5. fejezetei képezték, a magyar nyelvű szövegek esetében pedig – hasonló stílusú és nyelvhasználatú korpuszt keresvén – a Magyar Nemzeti Szövegtár Heti Világgazdaságot érintő részeit vizsgáltuk. Kísérleteink kitértek a szóalakok hasonlóságának meghatározásának különféle paraméterek melletti vizsgálatára: a kontextusablak mérete, akárcsak a hasonlósági gráf esetében a k legközelebbi szomszédság értékei 1 és 3 között mozogtak, továbbá megvizsgáltuk azt is, miképp befolyásolja a szóalakok csoportosításának eredményességet, ha eltérő nagyságrendű szöveg alapján hajtjuk végre mindazt. A két nyelvre elkészített eltérő nagyságrendű korpuszokkal kapcsolatos statsztikákat a 1. táblázat tartalmazza. (Mivel a Magyar Nemzeti Szövegtár esetében nem állt rendelkezésre az az információ, hogy egy szóalakra nézve melyek a szóba jöhető szófaji kódok, így ott a szóalakonkénti átlagos szófajszámot/többértelműséget nem állt módunkban kiszámolni.)

1. táblázat. Az angol és magyar nyelvű korpuszok statisztikái.

	WSJ		MNSZ	
	Szint_1	Szint_2	Szint_1	Szint_2
Mondatok száma	7053	34486	6069	30524
Tokenek száma	145002	723415	145006	723416
Szóalakok száma	13750	31686	36224	110133
Átlagos tokengyakoriság	10,55	22,83	4,00	6,57
Szóalakonkénti átlagos szófaj	$2.26 \pm 1{,}38$		-	

A nagyobb gráfok (Szint_2) esetében megvizsgáltuk a T-Man hálózatitopológia-közelítő algoritmus konvergenciájának sebességét az iterációk tükrében, ami az 1. ábrán látható. Az egyes iterációkhoz tartozó szaggatott vonalak alapján leolvasható, hogy átlagosan hány százalékkal haladta meg a közelített gráfokban szereplő élek összsúlya az etalon k-legközelebbi gráfok alapján elvárható összsúlyokat. A folytonos vonalak mentén az látható, hogy az egyes iterációk után a gráf csúcsaihoz választott legközelebbi szomszédok mekkora hányada volt megtalálható a tényleges – de csak jóval több számítás árán megkapható – k-legközelebbi szomszédságban szereplő élekhez képest. A körrel jelzett értékek a magyarra, a csillaggal jelzettek pedig az angol eredményekre vonatkoznak.

A felügyelet nélküli szófaji kódolás hatékonyságát jellemzően a kialakult klaszterek tényleges szófaji csoportokhoz való hozzárendelhetősége, valamint információelméleti szempontok szerint szokás vizsgálni. Eredményeink a megszo-

kott **V1-mérték**, illetve *'egy-az-egyhez'* (**1-1**) és *több-az-egyhez'* (**t-1**) értékek szerint kerülnek közlésre.

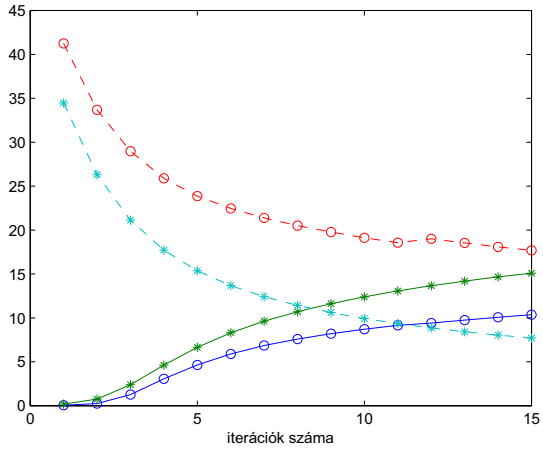
2. táblázat. A három fő paraméter (távolságszámítás módja, figyelembe veendő legközelebbi szomszédok száma, kontextusablak mérete) közül pontosan egy lefixálása mellett elért átlagos eredmények az eltérő méretű és nyelvű szövegeken.

	MNSZ						WSJ					
	$Szint_1$			$Szint_2$			$Szint_1$			$Szint_2$		
	V1	1-1	t-1	V1	1-1	t-1	V1	1-1	t-1	V1	1-1	t-1
COS	0.3336	0.2646	0.3929	0.3493	0.2793	0.4266	0.4466	0.3054	0.5501	0.4711	0.3150	0.5907
JS	0.3096	0.2260	0.3581	0.3345	0.2415	0.3800	0.4011	0.3034	0.4681	0.4631	0.3425	0.5343
JACC	0.2558	0.1880	0.2924	0.2799	0.2049	0.3142	0.3184	0.2446	0.3993	0.3204	0.2323	0.3960
k=1	0.4138	0.2510	0.4715	0.4322	0.2569	0.5212	0.4747	0.3115	0.6283	0.4932	0.3053	0.6803
k=2	0.2474	0.2164	0.2943	0.2726	0.2295	0.3013	0.3385	0.2640	0.3950	0.3875	0.3025	0.4339
k=3	0.2378	0.2111	0.2777	0.2589	0.2393	0.2982	0.3529	0.2778	0.3942	0.3740	0.2819	0.4068
w=1	0.3270	0.2316	0.3768	0.3281	0.2308	0.3838	0.3894	0.2702	0.4506	0.4258	0.2857	0.5137
w=2	0.2956	0.2342	0.3475	0.3275	0.2531	0.3820	0.3860	0.2964	0.4531	0.4380	0.3341	0.5317
w=3	0.2764	0.2127	0.3191	0.3083	0.2417	0.3549	0.3111	0.2498	0.3887	0.3909	0.26700	0.4755

3. táblázat. A nagyobb mennyiségű szövegekből készített k-legközelebbi szomszédsági gráf közelítő meghatározása segítségével elért átlagos eredmények pontosan egy paraméter lefixálása mellett.

	MNSZ			WSJ		
	V1	1-1	t-1	V1	1-1	t-1
COSINE'	0.3167	0.2645	0.3896	0.4724	0.3364	0.5859
JS'	0.2562	0.2052	0.3083	0.4029	0.2924	0.4720
JACC'	0.2135	0.1756	0.2665	0.2662	0.2090	0.3575
k'=1	0,3923	0,2494	0,4770	0,485	0,3073	0,6532
k'=2	0,2049	0,2009	0,2512	0,3399	0,2775	0,3946
k'=3	0,1883	0,1950	0,2363	0,3167	0,2530	0,3675
w'=1	0,2645	0,2087	0,3264	0,3649	0,2593	0,4632
w'=2	0,2645	0,2226	0,3248	0,4009	0,3038	0,4916
w'=3	0,2564	0,2140	0,3132	0,3758	0,2747	0,4605

A 'több-az-egyhez' kiértékelés olyan megengedő értéket határoz meg a szóalakok csoportosításához, amely a megtalált közösségeket olyan módon rendeli az etalon szófaji címkék által alkotott szóalakok csoportjaihoz, hogy a pontosság maximalizálva legyen. Ezzel szemben az 'egy-az-egyhez' kiértékelés megköveteli azt a feltételt, hogy a megtalált csoportok hozzárendelése az etalon csoportokhoz kizárólag olyan módon történhet, hogy egy etalon csoporthoz egy közösséget rendelhetünk. Jelen eredmények az 'egy-az-egyhez' hozzárendelés mohó módon



1. ábra. A k-szomszédsági gráfok pletykaalgoritmussal történő közelítésének konvergenciája a végrehajtott iterációk számának függvényében.

történő meghatározása mellett értendők (amely nem feltétlen egyezik meg a globálisan legjobb hozzárendelés értékével). Természetesen ez utóbbi kiértékelés jobban bünteti azokat a felbontásokat, amelyek az etalon szerint elvártnál jóval nagyobb számú csoportot eredményeznek.

Az információelméleti alapokon nyugvó V1-mérték [8] az egy klaszterezéshez tartozó *homogenitás* és *teljesség* értékekből számított súlyozott harmonikus átlagaként áll elő, hasonlóan az osztályozások jóságát jellemző F-mértékhez, ami a pontosság és a fedés értékeket ötvözi. A homogenitás feltételes entrópiát használva számszerűsíti, hogy a kialakuló egyes csoportok mennyire diverzek az etalon csoportokhoz képest. A teljesség számítása analóg módon történik, a különbség mindössze annyi, hogy ennek esetében az etalon címkék diverzitása kerül számszerűsítésre a megtalált klaszterek fényében. Egy tökéletes klaszterezés esetében az összes egy etalon csoportba tartozó elemet ugyanabban a megtalált klaszterben kell találjunk. Hasonlóan az F-mérték általánosításához, a V-mérték esetében is lehetőség nyílik annak két összetevőjének egymáshoz mért fontossága alapján meghatározni – $\beta = 1$ választástól különböző módokon is akár – egyéb V_β értékeket.

5. Diszkusszió

A hasonlósági gráfok segítségével leghatékonyabban a főnevek, igék, segédigék és számnevek csoportjait sikerült azonosítani: minden általunk használt módszer elfogadható mértékben azonosította őket. Ez különösen igaz a hónapnevekre és a különféle cégformák rövidített alakjaira (például *Co.* vagy *Ltd.*), hiszen ezekben az esetekben szemantikailag hasonló szavak kerültek egy csoportba. A

fenti szófajokkal szemben a legkeményebb diónak a határozószavak bizonyultak. A határozószavak elég vegyes csoportot alkotnak (morfológiai jegyekkel és mondatbeli pozícióval kevésbé megfoghatók), így megfelelő osztályba sorolásuk nehézséget jelentett mindegyik módszer számára. Érdekes módon a k legközelebbi szomszéd és a Jaccard-módszer is azonos gráfba helyezte az elöljárókat, névelőket és kötőszavakat, aminek az lehet a magyarázata, hogy hasonló környezetben fordulnak elő (például gyakran főnévi előtti pozícióban). Megjegyezzük ugyanakkor, hogy e szófajok elkülönítése problémásnak nevezhető az angol nyelvben [13]. A szomszédok számának meghatározásával és az ablakméretek rögzítésével kapcsolatban ugyanakkor azt találtuk, hogy a kisebb értékek bizonyultak hatásosabbnak, tehát elsődlegesen a szavak szűk környezete befolyásolta a csoportokba sorolást.

Az egyes módszerek összevetését tekintve a Jaccard-módszer bizonyult leghatékonyabbnak az *-ing*-es alakok (gerund) azonosításában. A k legközelebbi szomszéd módszer a melléknevek felismerésében nyújtott kitűnő eredményt, továbbá hatékonynak bizonyult az igeként és főnévként egyaránt szereplő szóalakok csoportosításában (pl. *decrease*). Szintén e módszer remekelt a névelemek osztályba sorolásában, különösen az ország- és nemzetiségnevek besorolása bizonyult sikeresnek. Ez arra utalhat, hogy e módszer a felügyelet nélküli szófaji egyértelműsítés mellett felügyelet nélküli szemantikai osztályozásra is feltehetőleg jól használható.

A közösségkereső eljárás során elnagyoltabb és részletesebb lexikai csoportok is létrejöttek. Angol nyelvre az elnagyoltabb csoportosítás esetében sikeresnek bizonyult a névmások, többes számú főnevek, tulajdonnevek és melléknevek kezelése, ugyanakkor az igei és főnévi szerepet egyaránt betölthető szóalakok is egy osztályba kerültek. Ugyanez mondható el az elöljárószavakra és határozószavakra is. Az angol nyelvű finomabb osztályozás során a szófaji osztályozáson túl szemantikai csoportok is megjelentek (például egy közösséget alkot a *TV*, *video*, *radio* szócsoport), de a helynevek osztályozása is jónak mondható. Mindemellett külön csoportokba kerültek az előbb még egy osztályba sorolt prepozíciók és névelők, determinánsok.

Magyar nyelvű kísérleteinkben a főnevek, számnevek és segédigék azonosítása volt a legeredményesebb, az igék és névutók felismerése valamivel nehezebb feladatnak bizonyult. Az angolhoz hasonlóan a funkciószavak (kötőszavak, névmások, névelők, határozószavak) itt is egy osztályba kerültek mindegyik módszer alkalmazásakor. Mindezt szintén a hasonló mondatbeli pozíció magyarázhatja: a vonatkozó névmások például a kötőszavakhoz hasonló viselkedést mutatnak. Módszereinket összehasonlítva azt találjuk, hogy a névelemek azonosításában a Jaccard-módszer felülmúlja a másik kettőt, különösen igaz ez a politikai pártokra és a személynevekre, vagyis itt is képes szemantikai alapú névelemcsoportok létrehozására.

A közösségkereső eljárás által létrehozott csoportok a magyarban kevésbé bizonyultak jónak, mint az angolban. Noha itt is megfigyelhetünk szemantikai alapú csoportosítást (hét napjai, hónapok) a részletesebb osztályozásban, általánosságban a számnevek felismerése érte el a legjobb eredményt. Érdekes

módon a főnevek és melléknevek gyakran kerültek egy csoportba, amit valószínűleg az magyarázhat, hogy a magyarban mindkét szóosztály hasonló toldalékokat vehet fel (többes szám jele, birtokos jel, esetragok).

Ha összevetjük az angolra és magyarra kapott eredményeinket, azt láthatjuk, hogy a felügyelet nélküli szófaji egyértelműsítés könnyebb feladat angolon, mint magyaron. Ezt természetesen a nyelvek közti eltérésekre vezethető vissza. Egyrészt az angolban nagyságrendekkel kevesebb szóalak tartozik egy lemmához, mint a magyarban (erre utal a lehetséges szófaji kódok száma is). Másrészt a magyarban jóval kisebb a többértelmű szóalakok (homonimák) száma, az angol ezzel szemben bővelkedik az ige/főnév/melléknév stb. szerepben egyaránt előforduló szavakban (pl. *present*). Mindebből az következik, hogy a magyarban több szóalak fordul elő, így ezek csoportosítása is nehezebb feladat. Harmadrészt az angol szórendje kötött, míg a magyar szórend a mondat információs szerkezetét tükrözi, ami azt jelenti, hogy az osztályozandó szó környezete sokkal változatosabb lehet, mint az angolban, vagyis nehezebb a kontextus felett általánosítani.

6. Összegzés

Ebben a munkában bemutattuk felügyelet nélküli szófaji egyértelműsítő módszerünket, mely közösségkeresésre épül. A szóalakok fölött értelmezett hasonlósági gráf költséges számítására való tekintettel az elosztott rendszerek területén az ún. overlay topológiák közelítésére korábban már sikeresen alkalmazott T-MAN algoritmust alkalmaztuk. Angol és magyar nyelvű eredményeink egyaránt azt igazolják, hogy sikerült átültetnünk a két különböző tudományos közösség által használt módszerek előnyeit a szófaji egyértelműsítés területére, azaz egy olyan feladatra nyújtottunk így megoldást, amelyet egy harmadik tudományos közösség tűzött ki céljául.

Köszönetnyilvánítás

A kutatás – részben – a MASZEKER és BELAMI kódnevű projektek keretében a Nemzeti Fejlesztési Ügynökség, illetve a TÁMOP-4.2.1/B-09/1/KONV-2010-0005 jelű projekt keretében az Európai Unió támogatásával, az Európai Regionális Fejlesztési Alap és az Európai Szociális Alap társfinanszírozásával valósult meg.

Hivatkozások

1. Halácsy, P., Kornai, A., Oravecz, C.: HunPos - an open source trigram tagger. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, Prague, Czech Republic, Association for Computational Linguistics (2007) 209–212
2. Jelasity, M., Montresor, A., Babaoglu, O.: T-man: Gossip-based fast overlay topology construction. Comput. Netw. **53** (2009) 2321–2339

3. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment **2008**(10) (2008) P10008+
4. Biemann, C.: Chinese whispers: an efficient graph clustering algorithm and its application to natural language processing problems. In: Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing. TextGraphs-1, Stroudsburg, PA, USA, Association for Computational Linguistics (2006) 73–80
5. Lamar, M., Maron, Y., Johnson, M., Bienenstock, E.: Svd and clustering for unsupervised pos tagging. In: Proceedings of the ACL 2010 Conference Short Papers. ACLShort '10, Stroudsburg, PA, USA, Association for Computational Linguistics (2010) 215–219
6. Gao, J., Johnson, M.: A comparison of Bayesian estimators for unsupervised Hidden Markov Model POS taggers. In: EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Morristown, NJ, USA, Association for Computational Linguistics (2008) 344–352
7. Van Gael, J., Vlachos, A., Ghahramani, Z.: The infinite HMM for unsupervised PoS tagging. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore, Association for Computational Linguistics (2009) 678–687
8. Rosenberg, A., Hirschberg, J.: V-measure: A conditional entropy-based external cluster evaluation measure. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). (2007) 410–420
9. Christodoulopoulos, C., Goldwater, S., Steedman, M.: Two decades of unsupervised POS induction: How far have we come? In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Cambridge, MA, Association for Computational Linguistics (2010) 575–584
10. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Physical Review E **69**(2) (2004) 026113+
11. Biemann, C.: Unsupervised part-of-speech tagging employing efficient graph clustering. In: Proceedings of the 21st International Conference on computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. COLING ACL '06, Stroudsburg, PA, USA, Association for Computational Linguistics (2006) 7–12
12. Brandes, U., Delling, D., Gaertler, M., Goerke, R., Hoefer, M., Nikoloski, Z., Wagner, D.: Maximizing modularity is hard. (2006)
13. Santorini, B.: Part-of-speech tagging guidelines for the penn treebank project. Technical report, Department of Computer and Information Science, University of Pennsylvania (1990)